ZDM Mathematics Education (2016) 48:29–40
DOI 10.1007/s11858-016-0765-0

ORIGINAL ARTICLE

Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations

Lena Schlesinger[1] · **Armin Jentsch**[1]

Accepted: 5 February 2016 / Published online: 19 February 2016

Abstract In this article, we analyze theoretical as well as methodological challenges in measuring instructional quality in mathematics classrooms by examining standardized observational instruments. At the beginning, we describe the results of a systematic literature review for determining subject-specific aspects measured in recent lesson studies in mathematics education. The main results are that there is little or no consistency in the conceptualization and nomination of subject-specific aspects. We therefore structured these different aspects along two perspectives, a mathematical perspective on mathematics educational quality of instruction as well as a pedagogical perspective. Furthermore, referring to the usage of these observational instruments in the field, in this paper we inquire into methodological challenges in measuring instructional quality in mathematics classrooms, e.g., the optimal number of raters and lessons to be observed. The results are twofold: on the one hand, there are recent studies that provide a useful answer to these questions. On the other hand, these results appear to be specific to the given data. Therefore, this problem seems to be unsolved so far.

Keywords Instructional quality · Methodological challenges · Classroom observations

✉ Armin Jentsch
armin.jentsch@uni-hamburg.de

Lena Schlesinger
lena.schlesinger@uni-hamburg.de

[1] University of Hamburg, Hamburg, Germany

1 Introduction

Within the last decade, research on both teachers' professional competencies and teachers' performance in the classroom has been of major interest in mathematics education (Baumert et al., 2010; Charalambous, & Hill, 2012; Hill et al., 2008; Learning Mathematics for Teaching Project, 2011; Schmidt et al., 2007). As these characteristics have a long tradition of being polarized, a great step forward in the field of educational research has been the framework of Blömeke, Gustafsson, and Shavelson (2015). In this framework, competence is conceptualized as a continuum including teachers' dispositions, their situation-specific skills and their performance, namely the observable behavior in real classroom situations, with the situation-specific skills comprising the competence facets *Perception*, *Interpretation* and *Decision-making* (in short PID-model).

Recent studies focused mainly on the relation between teacher competence and students' achievements, analyzing to what extent competence is directly predictive of students' outcomes. With regard to the processes between these characteristics, instructional quality received increased attention as a variable mediating the relation between teachers' competence and students' achievements (Baumert et al., 2010; Kersting et al., 2012; Hill, Rowan, & Ball, 2005). The question of instructional quality or what good instruction *is* has a long history in educational research (e.g. Oser, Dick, & Patry, 1992). Depending on different research traditions, the approach to this question and possible answers vary. Recent foci emphasized methods of teaching, instructional goals, as well as learning theories, which resulted in educational reforms in several countries (Atweh, Clarkson, & Nebres, 2003; Matsumura, Garnier, Pascal, & Valdés, 2002; Sawada et al., 2002). Particularly in mathematics education and mathematics instruction "scholars, policymakers, and educators have spent decades debating what ‘counts’ in mathematics classrooms” (Learning Mathematics for Teaching Project, 2011, p. 25).

Quality of instruction is discussed as one of the main influential factors on students’ learning and achievement (Hattie, 2009; Hill et al., 2005). Seidel and Shavelson (2007) identified the largest teaching effects on students’ outcomes for domain-specific aspects of teaching. Hiebert and Grouws (2007) claimed that “the nature of classroom mathematics teaching significantly affects the nature and level of students’ learning” (p. 371). However, the most relevant subject-specific aspects of instructional quality in mathematics education have yet to be identified as these aspects are inconsistent in different studies. In addition, depending on learning goals, it is not unambiguous how mathematics teaching can be effective (Hiebert & Grouws, 2007).

Several instruments have recently been developed for measuring instructional quality; amongst others, in Germany this development was influenced by the unsatisfactory achievements of German students in international large scale assessments such as TIMSS or PISA (Beaton, Mullis, Martin, Gonzales, Kelly, & Smith, 1996). In addition, research on instructional quality has a long tradition in the U.S. resulting for example in a large number of instruments on the measurement of educational reforms (Learning Mathematics for Teaching Project, 2011; Smith & Gorard, 2007; Scheerens & Bosker, 1997). Overall, a major goal of these approaches is to describe teachers’ professional competence and classroom practice in order to improve teaching and teacher education (Scheerens, 2004).

However, most of these instruments, which have been developed by researchers and practitioners, do not consider subject-specific aspects of instructional quality as the instruments are used for different subjects, which means that subject-related aspects, for example concerning mathematics education, are not part of the evaluation.

Not unexpectedly, methodological challenges arise when measuring instructional quality reliably, particularly when regarding effects on students’ outcomes (Hill et al., 2010, 2012; Praetorius et al., 2014). However, these challenges have frequently not been discussed in empirical research publications so far, although research literature points out that influences of teacher competence on instructional quality and its effects on students’ achievement depend strongly on methodological considerations (e.g. Hiebert & Grouws, 2007; Praetorius et al., 2012).

The aim of this paper is therefore twofold: we focus on theoretical as well as methodological challenges in measuring instructional quality in mathematics education with classroom observations which means observations of lessons (videos as well as live observations) by external persons using standardized rating instruments. Firstly, the paper aims at conceptualizing instructional quality and discusses how to integrate subject-specific aspects into a generic model. For this purpose, we provide an overview of important recent instruments measuring instructional quality in mathematics education. Secondly, the paper examines methodological considerations in measuring instructional quality using observer ratings. These analyses finally lead to prospects concerning the future development of the measurement of instructional quality in mathematics education. Our aim is not to identify the “best” or “most adequate” instrument for measuring instructional quality. In contrast, we present theoretical and methodological challenges to researchers in mathematics education, for facilitating the selection of instruments measuring instructional quality in mathematics classrooms.

2 Conceptualization of instructional quality

In the last decades, *teaching effectiveness* research seems to be one of the most prominent strands in the fields of psychology and education (Brophy, 2000; Kersting et al., 2012; Oser et al., 1992). In this strand, instructional quality is understood in a more functional way, i.e., the main goal is to predict students’ achievements at school (Seidel & Shavelson, 2007). These studies are often based on the *process-mediation-product paradigm* (Brophy, 2000, 2006), which emphasizes relations between aspects of instruction as “opportunities to learn” provided by the teacher (process), students’ usage (mediation) and their achievement (product). This framework has been modified to *the utilization of learning opportunities model* (Fend, 1981; Helmke, 2012), which is based on the constructivist idea that students’ learning processes cannot be controlled from outside. The teacher’s task is to provide learning opportunities that have to be used effectively by students in order to develop their achievements. Scholars conducting empirical research tend to describe instructional quality by a setting of characteristics independently from certain instructional designs. As recent studies on instructional quality have created many individual results it has become important to structure or “meta-analyze” the relevant factors of effective instruction (Seidel & Shavelson, 2007). However, these approaches were mainly *inductive*, i.e. they were not based on learning theories (Pianta & Hamre, 2009).

Recently, another framework for instructional quality has gained attention. It has been developed by several studies from German-speaking countries within the TIMSS Video Study and consists of three dimensions which are called *classroom management*, *personal learning support* and *cognitive activation* (see Fig. 1). It treats instructional quality in more detail, but separately

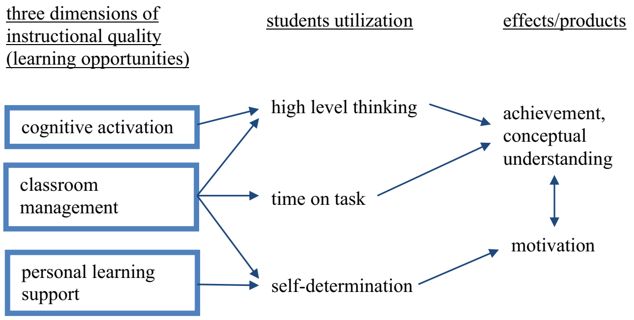

Fig. 1 Three-dimensional model of instructional quality (reproduced from Klieme & Rakoczy, 2008, p. 228)

from other facets like subject, school and grade even though it was firstly developed for mathematics instruction (Baumert et al., 2010; Lipowsky et al., 2009; Lotz, Lipowsky, & Faust, 2013; Helmke, 2012). Several studies have used this or a similar three-dimensional framework as a foundation for further empirical research (Klieme, Pauli, & Reusser, 2009; Pianta & Hamre, 2009). The predictive validity of these three basic dimensions on students' achievement has already been pointed out (Baumert et al., 2010; Klieme et al., 2009; Kunter et al., 2013; Lipowsky et al., 2009).

The first dimension, *classroom management*, focuses on quality-oriented learning time provided for students. Amongst other aspects, this dimension focuses on how effectively the teacher deals with disruptions or disciplinary conflicts (Brophy, 2000; Kunter, Baumert, & Köller, 2007). Effective *classroom management* is characterized by a structured and well-organized lesson with clear rules and routines (Lipowsky et al., 2009; see also Kounin, 1970).

The second dimension, called *personal learning support*, focuses on aspects of self-determination theory (Deci & Ryan, 1985). This dimension includes students' individual support provided by differentiation, the creation of a positive learning climate with a good relationship between students and teacher as well as constructive feedback (Klieme et al., 2009; Lipowsky et al., 2009).

Finally, *cognitive activation* refers to a high level of students' thinking supported by the teacher with problem-solving tasks to activate learning and understanding processes (Brophy, 2000; Hiebert & Grouws, 2007; Klieme et al., 2009; Lipowsky et al., 2009). This dimension includes the activation of previous knowledge, the activity of co-construction beginning with student ideas as well as challenging tasks and questions (Lipowsky et al., 2009; Praetorius et al., 2014). With regard to subject-specific aspects of instructional quality in mathematics education, the conceptualization of the third dimension can differ largely between studies (see Sects. 2.1, 2.3).

2.1 Instructional quality: generic or subject-specific?

Since the three-dimensional framework we just described was developed to hold for many different subjects, it does not sufficiently describe the *mathematics educational* quality of instruction. It becomes obvious that the domain *content* is not completely covered (Klieme & Rakoczy, 2008; Praetorius et al., 2014). However, the importance of subject-specific aspects on instructional quality has been claimed widely from theoretical and empirical perspectives (Drollinger-Vetter, 2011; Hiebert & Grouws, 2007; Klieme & Rakoczy, 2008; Seidel & Shavelson, 2007).

Despite these claims, the *relation* between subject-specific and generic aspects of instructional quality has been treated seldom in mathematics education and it is not clear until now to what extent the generic dimensions include subject-specific aspects especially regarding *cognitive activation*. Furthermore, the dimensionality of subject-specific aspects has rarely been addressed in mathematics education. Various studies aimed at solving this problem by either integrating a few subject-specific aspects into the three basic dimensions of instructional quality or by creating additional dimensions. We explore these approaches first before proposing a more systematic way of including subject-specific aspects in educational quality of instruction in Sect. 2.3.

Whether a dimension could be understood as subject-related depends essentially on its operationalization (Baumert et al. 2010; Drollinger-Vetter 2011). Educational researchers have mostly agreed on what they mean by *classroom management*, whereas other dimensions of instructional quality are not as well-established (Praetorius et al., 2014). *Classroom management* is mainly regarded as a generic dimension, because the aspects of this dimension are important in all subjects (e.g., clear rules and routines and well-organized lessons, see description in Sect. 2). Still, it is possible that reasons for students' disruptions are partly content-related due to domain-specific interests or content-specific mental under- or overload (Drollinger-Vetter, 2011, p. 325). Furthermore, there are different kinds of operationalization of *structural clarity*. Several researchers classify *structural clarity* as a generic sub-dimension of *classroom management* focusing on *organizational* structure and clarity (e.g. Helmke, 2012; Kunter et al., 2007). In contrast, other researchers regard *structural clarity* as a subject-specific dimension focusing amongst others on the development and implementation of mathematical concepts. As an example, in the German-Swiss project *Instructional Quality and Mathematical Understanding in Different Cultures* (so-called "Pythagoras-study"; Klieme et al., 2009) subject-specific aspects were measured separately beyond the three-dimensional generic framework. This subject-specific dimension consists of elements of understanding, the quality of representations and structural clarity in a content-related operationalization (Drollinger-Vetter & Lipowsky, 2006; Drollinger-Vetter, 2011, p. 179). However, it remains uncertain whether all authors who use the term "structural clarity" refer to the same kind of structure.

The dimension *personal learning support* can also be regarded as generic, focusing on positive learning climate and the relationship between students and teacher (Klieme & Rakoczy, 2008; Pianta & Hamre, 2009). Still, various kinds of operationalization differ between studies as other researchers include content-specific activities that support students' learning (Baumert et al., 2010). In mathematics classes, subject-specific aspects of *personal learning support* could be content-related adaptive support, inner differentiation based on different content foci, as well as a positive approach to students' conceptual errors or misunderstandings.

Especially the operationalization of *cognitive activation* differs largely between various studies and it is ambiguous, to what extent *cognitive activation* contains subject-specific aspects. Beyond the approach to use "challenging tasks" or foster "high-level thinking", it is still not clear what such a dimension could consist of. Several researchers describe *cognitive activation* by aspects like "activating previous knowledge", "building on students' ideas" or "stimulating students to explain their solution methods" (Klieme et al., 2009), whereas others include aspects like "scaffolding" (Schoenfeld, 2013). Especially when focusing on students' mathematical concepts and solution methods, these aspects are mainly subject-specific. However, aspects like "fostering high-level thinking" or "using scaffolding" might occur not only in mathematics education, but in other subjects too and can therefore be described as generic. For instance, researchers in the German study *Professional Competence of Teachers*, *Cognitively Activating Instruction*, *and Development of Students´ Mathematical Literacy* (in short COACTIV-study; Baumert et al., 2010) developed a framework according to the three basic dimensions of instructional quality with a focus on the *potential for cognitive activation*. By doing so, *cognitive activation* was understood from a content-specific perspective with a focus and an emphasis on cognitively activating and mathematically demanding tasks (Baumert et al., 2010; Klieme et al., 2009; Lipowsky et al., 2009).

So, even if one might use the same framework and the same dimensions of instructional quality, it is possible that the interpretation of a certain dimension and the subject-specific depth may vary greatly. This leads to several empirical problems as analyses between instructional quality and students' achievement can differ largely between different studies due to known conceptual differences (Seidel & Shavelson, 2007).

Overall, there is a shortcoming of frameworks that consist only of generic dimensions, namely, they do not seem to suffice as a theoretical framework to describe mathematics instruction completely. Even if several aspects of instructional quality are partly subject-specific, be it *cognitive activation* or *structural clarity*, many important details of teaching and learning are not addressed, such as language, representations, or correctness of the results. With regard to mathematics education and instructional quality, Blum, Drüke-Noe, Hartung, and Köller (2006) therefore refer to an educational approach describing mathematics education as orchestration of rich subject-related content beyond the three generic dimensions. In the following we use *mathematics educational quality of instruction* (MEQI) to describe those aspects of instructional quality that are specific to mathematics as a subject.

2.2 Subject-specific aspects of instructional quality

Conceptualizing subject-specific aspects of instructional quality seems yet to be a major theoretical challenge in educational research: "if educators could more satisfactorily describe and measure the MQI [mathematical quality of instruction], they would be in a better position to improve teaching and learning" (Learning Mathematics for Teaching Project 2011, p. 30).

In the following, we shall concentrate on subject-specific aspects of instructional quality in detail and how they have been addressed by recent studies. In order to gain a representative insight into the field we first conducted an explorative survey and looked at a wide range of different instruments that measure instructional quality, although most of these instruments do not focus on subject-specific aspects of instructional quality. For validating our findings we then conducted a systematic literature review within the international databases of Web of Science, ERA and ERIC. In the description of our research we included only articles from social sciences, educational sciences or psychology. Our selection criterion was the publication source of the papers or the instruments, restricting ourselves to papers coming from peer-reviewed journals that have been published within the last 20 years. The following keywords were used within the selection process: *instructional quality*, *quality of instruction*, *teaching quality*, *educational quality.* We crossed each of these keywords with the following terms: *mathematics*, *mathematics education* or *mathematics instruction*. By reviewing the titles and abstracts, we then excluded articles that focus on a specific part of educational science, i.e., distance education, special education or preschool education. Finally, we excluded papers that do not use classroom observations to assess instructional quality.

In the following table we have not analyzed a given paper itself; the focus was on the standardized observational instrument that is described in this paper. As an example, there has apparently been more than one publication on the

MQI instrument, but in order to provide clarity we have listed the MQI instrument referring only to one publication instead of listing every publication in which the instrument is described. We want once again to highlight that the table does not include classroom observation instruments that do not focus on subject-specific aspects in mathematics education, as for example the CLASS instrument (Pianta & Hamre, 2009). Moreover, due to unpublished descriptions of potential subject-specific aspects it was not possible to find examples for each dimension listed in the table. Generic dimensions are not included in the table, as well as the development of the presented instruments; we are focusing only on the final conceptualizations.

We are aware of the fact that we might have missed important contributions in our study; however, this exploratory approach will give an overview of studies that have been carried out in this field. In order to get an overview of different instruments measuring instructional quality, we included the results of other research in mathematics education based on literature reviews (Learning Mathematics for Teaching, 2011; Schoenfeld, 2013). As in our case, the authors of these publications, too, could find only a small number of studies and instruments.

When looking at the different subject-specific aspects or dimensions, it becomes obvious that there is no consistency in either the nomination or the conceptualization of subject-specific aspects. In addition, it is confusing that some studies use the term *dimension* whereas other studies talk only about *aspects* or even *items*. Therefore, one aspect of a given study could cover a whole dimension within another study which makes it very difficult to compare the instruments described in Table 1. Even more generally, the dimensionality of subject-specific aspects has rarely been studied empirically in mathematics education. In addition, it shows that the relevant aspects of mathematics educational quality have yet to be identified, not to mention the relation between those aspects. On the other hand, it becomes apparent that there are some aspects that are included in many of the above listed instruments (e.g. representations, demanding tasks).

Therefore, a first approach to identify these commonalities can be to find out the aspects that are included in more than one instrument. The aspects covered are:

- Representations
- Mathematical language
- Mathematical content and topics (e.g. problem solving, reasoning)
- Connections, relations, abstractions and generalizations (i.e. mathematical richness)
- Mathematical errors, mathematical correctness
- Elements of understanding
- Instructional practices (classroom practices)
- Implementation of the task
- Students' participation and understanding
- Cognitive demand, cognitive activation, potential of the task
- Materials, manipulatives.

When analyzing these different conceptualizations, it is also important to consider that some instruments were developed for special content areas (in particular introduction to the Pythagorean Theorem), whereas others were not even developed for mathematics instruction solely. We will discuss this fact in Sect. 3.2 when describing methodological challenges. In the following, we develop some suggestions of how the subject-specific aspects in mathematics education could be structured in a more systematic way.

2.3 Structuring generic and subject-specific aspects of instructional quality

Overall, we can state that there is presumably more than one way to identify categories for mathematics educational quality of instruction. However, as we lack theories on the subject-specific depth of instructional quality (as mentioned in Sect. 2.1) we might as well move forward with a rather pragmatic approach. When we ignore the *terms* that were used to name subject-specific aspects in those studies (second column of Table 1), but look mainly at the content of a given aspect (see the examples), we see at least two independent approaches towards a mathematics educational quality of instruction.

Firstly, we find in the studies the attempt to define those facets of instruction that are subject-specific in a narrow way, for instance *mathematical language*, *mathematical errors or mathematical concepts*, *topics and connections*. By these aspects, one addresses aspects that are not relevant only in classroom practice but also in other situations that deal with mathematics; that is, these aspects might come from a more *mathematical perspective* on mathematics educational quality of instruction. In most cases, it is not viable to study *mathematical concepts* in other than mathematics classes, in contrast to the issue of *classroom practice* or *cognitive activation*, which can be studied in all kinds of classes. Consequently, we argue that these genuine subject-specific aspects of mathematical instruction cannot be integrated into a generic framework that aims at measuring instructional quality in any classroom.

Secondly, subject-specific dimensions can be found which arise when looking at mathematics educational quality of instruction from a *pedagogical perspective*. This can be done either theoretically or in a way that is related to educational practice. In this category, we can put subject-specific aspects like *instructional practices*, *connecting classroom practice to mathematics* or *students' discussions*

Table 1 Overview of subject-specific aspects of instructional quality measured in different studies

Study	Subject-specific dimensions/aspects (*examples*)
Elementary mathematics classroom observation form (Thompson & Davis, 2014)	Computation and concepts (*problem-solving*, *reasoning*, *algebra*, *geometry*, *measurement*, *data*,…) Technology (*calculators*, *computers*, *others*) Manipulatives (*materials*, *real-world objects*, *journals*, *pictures*, *textbooks*)
Inside the classroom observation protocol (Horizon Research, Inc., 2000)	Mathematics content (*significance*, *appropriateness*, *content information*, *concepts*, *abstractions*, *connections*, *sense-making*)
Instructional quality assessment (IQA) (Matsumura et al., 2002)	Potential of the task Implementation of the task Student discussion following the task Rigor of expectations
Mathematical quality of instruction (MQI) (Learning Mathematics for Teaching Project, 2011)	Richness of the mathematics (*representations*, *explanations*, *mathematical language*, *multiple procedures*, *developing generalizations*) Mathematical errors and imprecisions (*major mathematical errors*, *imprecision in language*, *lack of clarity*) Students participating in meaning making and reasoning (*providing explanations*, *posing mathematically motivated questions*, *engaging in reasoning*) Working with students and mathematics Connecting classroom practice to mathematics
PERLE (Lotz, Lipowsky, & Faust, 2013)	Elements of understanding Mathematical language Representations
Pythagoras study (Klieme et al., 2009)	Elements of understanding (*occurrence*, *duration*) Representations (*enactive*, *iconic*, *formal*, *verbal*) Structural clarity
Reformed teaching observation protocol (RTOP) (Sawada et al., 2002)	Propositional knowledge (*conceptual understanding*, *fundamental concepts*, *abstractions*, *connections*) Procedural knowledge (*representations*, *reflection*, *intellectual rigor*, *thought-provoking activity*)
TIMSS video study 1999 (Hiebert et al., 2003)	Instructional practices (*solution methods*, *mathematical processes*, *applications*, *problem context*) Mathematical content (*topics*, *reasoning*, *complexity*, *mathematical relations*)
TRU Math (Schoenfeld, 2013)	Mathematical focus, coherence and accuracy (*richness and centrality*, *mathematical practices*,…) Cognitive demand Equitable access to content
Uteach teacher observation protocol (UTOP) (Marder, & Walkington, 2014)	Lesson structure (*organization*, *important concepts*, *students' understanding*) Implementation (*problem-based approach*, *conceptual understanding*) Mathematics content (*deep knowledge and fluidity of the teacher*, *abstractions*, *connections*)
Capturing teacher knowledge (Kersting et al., 2012)	Developing concepts (*mathematical concepts or ideas are mathematically correct*) Appropriate use of representations to explain algorithms (*manipulatives and drawn representations*) Connecting concepts and topics

and *students*' *participation.* Moreover, there are facets of mathematics educational quality of instruction that are strongly related to the three generic dimensions of instructional quality. Apparently, under this category falls the *potential of the task* from IQA, *cognitive demand* from TRU Math and *cognitive activation* from PERLE and the Pythagoras study. We can identify particular learning theories (e.g., social constructivism) that have apparently been the conceptual foundation for these quality dimensions.

This second, pedagogical perspective of subject-specific aspects is, in our opinion, the more complicated part: due to the pedagogical perspective on mathematics instruction some aspects may be seen as generic. This is of course only partly true, as we have discussed earlier using the examples of *cognitive activation* and *structural clarity*. Thus, we can understand these mathematics educational aspects of instructional quality as both subject-specific and also generic (or in-between). Instructional quality can be seen in a direct connection to teachers' professional knowledge, as it is then interpreted as teachers' performance in the classroom (Blömeke et al., 2015; see also Kersting et al., 2012). It may hence be possible that instructional quality can be divided into separate subject-specific dimensions, similarly to the well-known division of teachers' knowledge into MCK, PCK and GPK. Broad discussions exist especially on the operationalization of PCK (see e.g. Buchholtz, Kaiser, Blömeke, 2014), which is supposed to have a major impact on students' learning (Shulman, 1986, 1987). Nevertheless, the conceptual issues we just mentioned can be regarded only as a first step towards future developments.

3 Methods and methodological challenges for measuring instructional quality

Beyond the question of how to conceptualize instructional quality, methodological challenges should also be addressed more carefully in order to gain a better understanding of how the research design affects the psychometric quality of a study (Hill, Charalambous, & Kraft, 2012; Hill, Kapitula, & Umland, 2010). The following table gives an overview of methods and methodological issues from several studies in which the instruments from Sect. 2.2 were used (Table 1). It becomes apparent that, regarding methodological decisions, these studies vary largely.

Classroom observation is the most commonly used method to measure instructional quality directly (Clare, Valdés, Pascal, & Steinberg, 2001; Helmke, 2012).[1] These observations can be conducted by either *internal* or *external* observers (or both). Internal observers are teachers or students who attend the class. According to many authors, internal observers have some disadvantages (e.g. Hiebert & Grouws, 2007; Pianta & Hamre, 2009): as teachers and students participate actively in the lesson, they cannot concentrate exclusively on the observation. In addition, self-reports by teachers are influenced by how teachers think they should be teaching. Hence, classroom observations by external observers are awarded higher validity. In addition, external observers generally look at a huge number of lessons of different teachers, so their ratings tend to be less subjective than the evaluation by teachers or students (Praetorius et al., 2012; Helmke, 2012).

However, there are a few well-known disadvantages that go along with external ratings (Howard et al., 1980; Lüdtke et al., 2009; Reyes et al., 2012). External observers spend little time on the observation compared with the long-term perspective of the teacher or students. In addition, the observers normally have only a little information about the class and the students. Nevertheless, these disadvantages are seen as less serious than those that go along with the evaluation by teachers or students (Clare et al., 2001; Helmke, 2012; Pianta & Hamre, 2009; Praetorius et al., 2014). As an example, Kunter and Baumert (2006) claim that there are only a few aspects of instructional quality that can be observed by teachers and students adequately, for example sampling information about the frequency of different instructional practices in a classroom or students' perceptions regarding the classroom climate or learning environment (see also Praetorius et al., 2014).

3.1 Assessing instructional quality with high-inferent observer ratings

Classroom instruction can be divided into aspects that can be observed directly (so-called *surface structures*, e.g., how many students put their hands up) and aspects that need interpretations (so-called *deep structures*, e.g., how many students participate in high-level thinking), which accounts for the amount of *inference* that is left to the observer. When analyzing an instrument, one can distinguish between items of higher or lower inference, i.e., the aspect to be measured requires more or less interpretation respectively, from the observer (Rosenshine, 1970). Nowadays, instructional quality is described mainly as a latent construct requiring interpretations, because recent research showed that the presence of aspects from surface structures and the quality of the deep structures can vary almost independently from each other (Baumert et al., 2010; Veenman, Kenter & Post, 2000). Hence, a questionnaire that contains only low-inference-items will lead to validity problems, even if the instrument is reliable. Of course, items with less inference are more likely to measure reliably and, therefore, have frequently been recommended in the past (Brophy, 2006; Kounin, 1970; Soar, Medley, & Coker, 1983). However, the main goal in measuring instructional quality must be that the ratings are not only reliable, but also valid (Hill et al., 2010; Kane, 2006; American Educational Research Association/American Psychological Association, 1999). Still, there are also some facets of instructional quality with lower inference, which can hence be evaluated more easily (e.g., students' oral participation in class vs. high-level thinking). In some cases, the indicators for the surface structure are an indication for the quality at deep

[1] Sometimes the word instructional research is used only if classroom observation methods are performed, as for example Helmke puts it: "The silver bullet of the description and assessment of instruction is without doubt observation" (own translation, 2012, p. 288).

structure; for example, mathematical high-level thinking of the students can only occur when the lesson time is connected to the learning of mathematics (i.e. time-on-task). For describing both surface and deep structures of mathematics teaching, an instrument is necessary that contains items at any inference level. The studies presented in Table 2 have measured instructional quality mainly with items of both lower and higher inference.

3.2 Reliability issues of external observer ratings

At a methodological level, the reliability of classroom observations—that is a necessary condition for valid measures—has been a major issue in the last few years (Hill et al., 2010, 2012; Praetorius et al., 2014). More precisely, the following questions are discussed in detail:

- How many lessons are sufficient in order to secure reliable data?
- How many raters are necessary?
- What is the optimal length of a rating period/time interval (*i.e. unit of analysis*)?
- How can different lessons be compared objectively?

In order to analyze these questions, the generalizability theory or, in short, G Theory is a useful psychometric method (for more information see Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). With G Theory it is possible to investigate different sources of error that can occur in classroom situations (Brennan, 2001; Shavelson & Webb, 1991; Hill et al., 2012). For every potential source of error, variance estimations are performed. The resulting G coefficients can be interpreted as reliability coefficients of classical test theory. Decision studies can then be conducted for estimating the reliability under different conditions (e.g. number of lessons and raters per class) and deciding on the optimal conditions for obtaining sufficiently reliable measures. With regard to other theories, reliability issues as mentioned above can be addressed less conditionally than in item response theory and more precisely than in classical test theory, which makes G Theory a great tool for analyzing instruments (Brennan, 2011; Hill et al., 2012). In conclusion, the decision studies enable researchers to address issues such as *rater bias*, i.e., errors caused by observers. Studies showed that errors committed by observers sum up to 41 % of the variance (Praetorius et al., 2012, 2014; for more details see Shavelson & Webb, 1991). However, in most cases, these analyses are limited to a single instrument since it is usually not possible to compare questionnaires validly, because of different concepts or terms (Hill et al., 2010; Seidel & Shavelson, 2007).

In the following, we present selected recent results on these questions and issues. The first question focuses on the variation of instructional quality between lessons and is asking, how many lessons are necessary to obtain sufficiently reliable measures for instructional quality. Until recently in empirical research, this issue has been discussed controversially by different researchers using plausibility arguments for both short periods and longer observation periods without empirical evidence. Praetorius et al. (2014) analyze this problem with data from the Pythagoras study using G theory. The results of their analysis show that the two quality dimensions *classroom management* and *personal learning support* (operationalized as content-independent) show a high stability across lessons, whereas the dimension *cognitive activation* (operationalized partly content-dependent) varied largely between lessons (Praetorius et al., 2014). In detail, one lesson is sufficient to measure *classroom management* and *personal learning support*, whereas nine lessons are needed to measure *cognitive activation* reliably. In addition, results of the MET-study suggest the benefit of observing each teacher during more than one single lesson (Gates Foundation, 2012). However, the questions remain whether these results are generalizable for further studies and how the operationalization (content-dependent vs. content-independent) influences the results.

The second question on the number of raters is discussed for the Pythagoras study by Praetorius et al. (2014). For reliability values greater than $E^2 = .90$ there are theoretically eight observers necessary for the evaluation of the dimension *personal learning support* (for more information see Brennan, 2001, 2011; Shavelson & Webb, 1991). Similar analyses were made for the MQI-instrument focusing on the number of lessons per teacher and the number of raters. The results show that the relatively small numbers of three lessons and two raters seem to be the optimal combination for their research purposes (Hill et al., 2012). Other studies, which were not presented here, found different optimal combinations that in some cases vary largely between the number of raters and lessons, characterized by measurement points (see Praetorius et al., 2014). However, these analyses depend strongly on the study within which observer reliability and rater bias have been measured and therefore a transfer to other studies is not easily possible. As Hill et al. (2012) put it: “there is no optimal number of observations or raters that transcends specific instruments and rater populations and we caution against extrapolating our results to other observational instruments and scoring designs” (p. 62). Even if no optimal number of raters generally exists, for analyzing rater bias, it is necessary that a minimum of two raters observe one lesson.

The third question addressing rating periods is another important methodological issue, which treats the *unit of analysis*. With regard to the relationship between

Table 2 Overview of methods and methodological issues for several studies measuring instructional quality in mathematics education

Study/instrument	Grades/classes	Number of lessons and ratings per lesson	Raters	Standardization of the lessons	Reliability
Algebra teaching study/TRU Math (Schoenfeld, 2013)	Grade 6–10 United States	10-min episodes	n.a.	No standardization	n.a.
Capturing teacher knowledge (Kersting et al., 2012)	Grade 5–7 36 teachers United States	1 lesson, 1 rating per lesson segment (several segments per lesson)	2 (and more)	Introduction to new fraction concept or idea	Matching percentage among raters: >.80
Elementary mathematics classroom observation (Thompson & Davis, 2014)	Grade 1–5 250 primary teachers Over 2000 observations United States	Two 15-min observation or 1 h per week, 24 weeks	n.a.	No standardization	n.a.
Horizon research/inside the classroom observation protocol (Horizon Research, 2000)	364 lessons United States	1 rating per lesson	n.a.	No standardization	n.a.
IQA (Matsumura et al., 2008)	Grade 6–7 13 teachers United States	2 observations, 1 rating per observation	1	Problem-solving activity; related discussions	Rater agreement (for 4 raters and 7 observations): $.11 < ICC \leq 1.0$
LMT project/MQI (Hill et al., 2012)	No specific grades Eight middle school teachers United States	6 lessons, 1 rating every 7.5 min (6–8 segments per lesson)	2	No standardization	Generalizability coefficient (for 2 raters and 3 lessons per teacher): $.68 < E\rho^2 \leq .81$
TIMSS video study 1999 (Hiebert et al., 2003/Jacobs et al., 2003)	Grade 8 638 classes Seven countries	1 lesson, 1 rating per lesson segment (several segments per lesson)	2	No standardization	Split-half: $.87 < r \leq 1.0$
PERLE video study mathematics (Lotz, Lipowsky, & Faust, 2013)	Grade 1–2 36 classes Germany	2 lessons, 1 rating per lesson	2	Introduction to multiplication	Rater agreement: $.60 < \kappa \leq 1.0$ Generalizability coefficient: $.70 < E\rho^2 \leq .99$
Pythagoras study (Klieme et al., 2009)	Grade 8–9 37 classes Switzerland and Germany	5 lessons, 1 rating per lesson	2 (3)	Introduction to Pythagorean theorem	Generalizability coefficient: $.57 < E\rho^2 \leq .88$ Internal consistency: $\alpha = .88$
ACEPT/reformed teaching observation protocol (Sawada et al., 2002)	153 classes Schools, colleges and university United States	3 lessons, 1 rating per lesson	1–2	No standardization	Rater agreement: *Pearson's* $r = .98$ Internal consistency: $.80 < \alpha \leq .93$
MET-study/Uteach teacher observation protocol (Marder & Walkington, 2014; Gates Foundation, 2012)	Grade 4–8 249 teachers 982 videos United States	2–4 lessons, 1 rating per lesson	1–3	No standardization	Reliability measure: $.44 < ICC \leq .65$

instructional quality and students' achievement it is important to analyze meaningful units that allow descriptions of the processes in class. Hiebert and Grouws (2007) proposed the analysis of units of typical lesson periods. Still, this decision depends on the aspects that are to be measured as there can be large variance in some aspects of instructional quality within one lesson, e.g., whether the lesson time is used efficiently. Therefore, recent observational instruments use shorter and therefore more rating periods (Learning Mathematics for Teaching Project 2011). This leads to other technical challenges, e.g., how many ad-hoc ratings can be performed per lesson if no videos are recorded.

The last question focuses on the comparability of lessons. In order to gain higher comparability, it is possible to standardize content, classes, grades, teacher activities, settings or learning goals within the lesson. Domain-specific teaching studies are mainly conducted under quasi-experimental conditions (Seidel & Shavelson, 2007). This quasi-experimental approach leads to some advantages and, therefore, is used in different recent studies like the Pythagoras study or the PERLE study. In both of these studies the content was standardized, e.g., all classes in the Pythagoras study treated an introduction into the Pythagorean Theorem. The instruments used in standardized content settings allow the deepening of the subject-specific perspective (Lipowsky et al., 2009) and the specification of the subject-specific items. However, this approach may have an influence on other methodological questions; for example, the question remains of how results of these studies can be transferred to other mathematical content and also, more globally, how these results can be transferred to ordinary all-day classroom instruction without any standardization. Therefore, one main goal in future research might be to develop instruments that can be applied in all-day instruction and that are not bound to specific contents or classroom situations.

4 Summary and conclusion

Until now, there seems to be no general framework for instructional quality, especially for the purposes of mathematics education and subject-specific aspects. Moreover, the understanding of what MEQI really *is* varies greatly among studies. This shows that there is apparently no consensus about the essential issues of MEQI, its structure and its relation to generic aspects. In addition, analyzing results of classroom studies with a focus on mathematics education becomes quite difficult as almost every framework uses different terms describing different underlying concepts. Therefore, we argue that agreeing on some definitions could be helpful for further research in mathematics education and, more generally, empirical research. As a first step, we suggest that MEQI could be studied from both a mathematical and a pedagogical perspective.

At the methodological level, we gave an overview of challenges concerning observer ratings. In detail, we focused on questions such as how many lessons and raters are necessary for securing reliable data and what is the optimal length of a rating period (unit of analysis). In addition, we discussed advantages and disadvantages that go along with quasi-experimental studies standardizing aspects of the observed lessons (e.g., the content). Still, there are other questions which could not be discussed in this paper (e.g., how raters can be trained ideally or the number of items in the observational instrument; see Hill et al., 2012; Praetorius et al., 2012). Consequently, Hill et al., (2012) use the term *observational systems* to highlight the interdependency of raters, instruments and conceptualizations in a given study. However, it is necessary also to think of systematic problems. In particular, one should address not only reliability, but even more important validity issues of instructional quality.

From the perspective of educational practice, the instruments described in Sects. 2.2 and 3.1 carry the problem of not always being easily applicable to large samples. They often require videotaping and, in order to analyze the instructional quality, need multiple inspections of videos, which leads to many organizational challenges (for more information see Casabianca et al., 2013). Hence, for conducting empirical research in mathematics classrooms it may be necessary to develop an observational instrument that can be used in ad-hoc classroom observations. Regarding cost-efficiency, statistical methods such as G theory can provide a useful benefit for practitioners as it is possible to decide on a theoretical basis how many raters and items may be necessary to gain reliable measures (see Hill et al., 2012; Shavelson & Webb, 1991). Thus, for future research it is important to use theory-based frameworks and observation instruments that fulfill both validity and reliability conditions for measuring instructional quality in mathematics education.

Acknowledgments We thank Nils Buchholtz, Andreas Busse and the reviewers for helpful suggestions and comments on earlier versions of this article.

References

American Educational Research Association/American Psychological Association. (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Atweh, B., Clarkson, P., & Nebres, B. (2003). Mathematics education in international and global contexts. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second* *international handbook of mathematics education* (pp. 185–229). Dordrecht: Springer Netherlands.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill: Boston College.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies. *Zeitschrift für Psychologie, 223*(1), 3–13.

Blum, W., Drücke-Noe, C., Hartung, R., & Köller, O. (2006). *Bildungsstandards Mathematik: Konkret. Sekundarstufe 1: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen*. Berlin: Cornelsen Scriptor.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. doi:10.1080/08957347.2011.532417.

Brophy, J. (2000). *Teaching*. Brüssel: International Academy of Education.

Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 755–780). Mahwah: Erlbaum.

Buchholtz, N., Kaiser, G., & Blömeke, S. (2014). Die Erhebung mathematikdidaktischen Wissens—Konzeptualisierung einer komplexen Domäne. *Journal für Mathematik-Didaktik, 35*(1), 101–128.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783.

Charalambous, C. Y., & Hill, H. C. (2012). Teacher knowledge, curriculum materials, and quality of instruction: Unpacking a complex relationship. *Journal of Curriculum Studies, 44*(4), 443–466.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools (CSE Technical Report No. 545)*. Los Angeles: National Center for Research on Evaluation.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability*. New York: Wiley.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior. Perspectives in social psychology*. New York: Plenum.

Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit: Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht*. Münster: Waxmann.

Drollinger-Vetter, B., & Lipowsky, F. (2006). Fachdidaktische Qualität der Theoriephasen. In E. Klieme, C. Pauli, & K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 3: Hugener, Isabelle; Pauli, Christine & Reusser, Kurt: Videoanalysen* (pp. 189–205). Frankfurt am Main: GFPF.

Fend, H. (1981). *Theorie der Schule* (2., durchges. Aufl). *U- & -S-Pädagogik*. München [u.a.]: Urban & Schwarzenberg.

Gates Foundation (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. Research paper*, http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf. Accessed 22 Jan 2016.

Hattie, J. (2009). *Visible learning. Synthesis of over 800 meta-analyzes relating to achievement*. London: Routledge.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.

Hiebert, J., Gallimore, R., Garnier, H., & Stigler, J. (2003). *Teaching mathematics in seven countries. Results from the TIMSS 1999 video study*. Washington: National Center for Education Statistics.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Charlotte: Information Age.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430–511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64. doi:10.3102/0013189X12437203.

Hill, H. C., Kapitula, L., & Umland, K. (2010). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794–831. doi:10.3102/0002831210387916.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406.

Horizon Research, Inc. (2000). Inside the classroom observation and analytic protocol. Chapel Hill: Horizon Research, Inc.

Howard, G. S., Maxwell, S. E., Weiner, R. L., Boynton, K. S., & Rooney, W. M. (1980). Is a behavioral measure the best estimate of behavioral parameters? Perhaps not. *Applied Psychological Measurement, 4*, 293–311.

Jacobs, J., Garnier, H., Gallimore, R., Hollingsworth, H., Givvin, K. B., Rust, K., Kawanaka, T., Smith, M., Wearne, D., Manaster, A., Etterbeek, W., Hiebert, J., Stigler, J. (2003). TIMSS 1999 video study technical report: volume 1: Mathematics study, NCES (2003-012), U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18–64). Westport: Praeger.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49*(3), 568–589. doi:10.3102/0002831212437853.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster: Waxmann.

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik, 54*, 222–237.

Kounin, J. S. (1970). *Disciplin and group management in classrooms*. New York: Holt, Rinehart and Winston.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231–251. doi:10.1007/s10984-006-9015-7.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*(5), 494–509. doi:10.1016/j.learninstruc.2007.09.002.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal* *of Educational Psychology, 105*(3), 805–820. doi:10.1037/a0032583.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education, 14*, 25–47.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527–537. doi:10.1016/j.learninstruc.2008.11.001.

Lotz, M., Lipowsky, F., Faust, G. (2013). *Dokumentation der Erhebungsinstrumente des Projekts "Persönlichkeits-und Lernentwicklung von Grundschülern"* (*PERLE*). *3. Technischer Bericht zu den PERLE-Videostudien. Materialien zur Bildungsforschung: Vol. 23,3*. Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung [u.a.].

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*, 120–131. doi:10.1016/j.cedpsych.2008.12.001.

Marder, M., & Walkington, C. (2014). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measuring Effective Teaching project* (pp. 234–277). New York: Wiley.

Matsumura, L. C., Garnier, H. E., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and students achievement. *Educational Assessment, 8*, 207–229.

Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment, 13*, 267–300.

Oser, F., Dick, A., & Patry, J.-L. (Eds.). (1992). *Effective and responsible teaching: The new synthesis*. San Francisco: Jossey Bass.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119. doi:10.3102/0013189X09332374.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction, 22*, 387–400.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*, 700–712. doi:10.1037/a0027268.

Rosenshine, B. (1970). Evaluation of instruction. *Review of Educational Research, 40*, 279–300.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253. doi:10.1111/j.1949-8594.2002.tb17883.

Scheerens, J. (2004). *Review of school and instructional effectiveness. Background paper prepared for the Education for All Global Monitoring Report 2005*. Paris: UNESCO.

Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.

Schmidt, W. H., Tatto, M. T., Bankov, K., Blömeke, S., Cedillo, T., Cogan, L., et al. (2007). *The preparation gap: Teacher education for middle school mathematics in six countries. Mathematics teaching in the 21st century (MT21)*. East Lansing: Michigan State University, Center for Research in Mathematics and Science Education.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM-The International Journal on Mathematics Education, 45*(4), 607–621.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. doi:10.3102/0034654307310317.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks: Sage.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–31.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Smith, E., & Gorard, S. (2007). Improving teacher quality: Lessons from America's No Child Left Behind. *Cambridge Journal of Education, 37*(2), 191–206.

Soar, R. S., Medley, D. M., & Coker, H. (1983). Teacher evaluation: A critique of currently used methods. *The Phi Delta Kappan, 65*, 239–246.

Thompson, C. J., & Davis, S. B. (2014). Classroom observation data and instruction in primary mathematics education: Improving design and rigour. *Mathematics Education Research Journal, 26*(2), 301–323. doi:10.1007/s13394-013-0099-y.

Veenman, S., Kenter, B., & Post, K. (2000). Cooperative learning in Dutch primary classrooms. *Educational Studies, 26*(3), 281–302.